

# CS 229br: Foundations of Deep Learning

## Lecture 3: Diffusion

Boaz Barak



Gustaf Ahdritz



Gal Kaplun

# Reminder 2

Exercise: If  $X \sim N(\mu, \Sigma)$  and  $Y = N(\mu', \Sigma')$  independent then  $X + Y \sim N(\mu + \mu', \Sigma + \Sigma')$

Cor: If  $X \sim N(\mu, \sigma^2 I)$  and  $Y \sim N(\mu', \sigma'^2 I)$  independent then  $X + Y \sim N(\mu + \mu', (\sigma^2 + \sigma'^2)I)$

Useful facts:

$$\mathbb{E}_{y \sim Y} H(X|Y = y)$$

$$I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

$$= \mathbb{E}_{y \sim Y} KL(X|Y = y \parallel X) = \mathbb{E}_{x \sim X} KL(Y|X = x \parallel Y)$$

$$(\log p(x))' = \frac{p(x)'}{p(x)} \quad \nabla p = p \cdot \nabla \log p$$

# Distances between Gaussians

Let  $X \sim N(\mu, \Sigma)$  and  $Y \sim N(\mu', \Sigma')$

If  $\Sigma = \Sigma' = I$ : Total variation:  $TV(X, Y) \approx 1 - \Pr[|N(0,1)| \geq \frac{1}{2} \|\mu - \mu'\|]$   
(If  $\|\mu - \mu'\| = \epsilon$  then  $TV(X, Y) \approx \epsilon$ )

$$KL : KL(X \parallel Y) = \frac{1}{2} \|\mu - \mu'\|^2$$

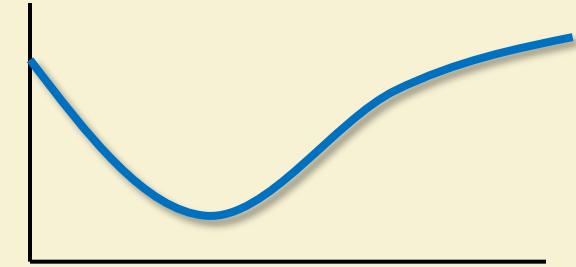
$$\text{Fr\'echet: } d^2(X, Y) = \min_{X', Y'} \|X' - Y'\|^2 = \|\mu - \mu'\|^2$$

$$\text{General: } KL: KL(X \parallel Y) = \frac{1}{2} \left[ \|\mu - \mu'\|_{\Sigma'^{-1}}^2 + \text{tr}(\Sigma'^{-1} \Sigma) - d + \log \frac{\det \Sigma'}{\det \Sigma} \right]$$

$$\text{Fr\'echet: } d^2(X, Y) = \|\mu - \mu'\|^2 + \text{tr}(\Sigma + \Sigma' - 2(\Sigma \Sigma')^{1/2})$$

# Reminder 3

If probability of  $x$  is proportional to  $\exp(-f(x))$



⇒ likely value of  $x$  is  $\min f(x)$

⇒ can find it by setting  $\nabla f(x) = 0$

# Useful fact

Let  $X, Y$  be random variables.

Let  $\mu(x) = \mathbb{E}[Y|X = x]$ . Then

$$\mu(x) = \arg \min \mathbb{E}[(y - Y)^2 | X = x]$$

Proof: Fix any  $x$  and let  $\mu = \mu(x) = \mathbb{E}[Y|X = x]$

Let  $f(y) = \mathbb{E}[(y - Y)^2 | X = x]$

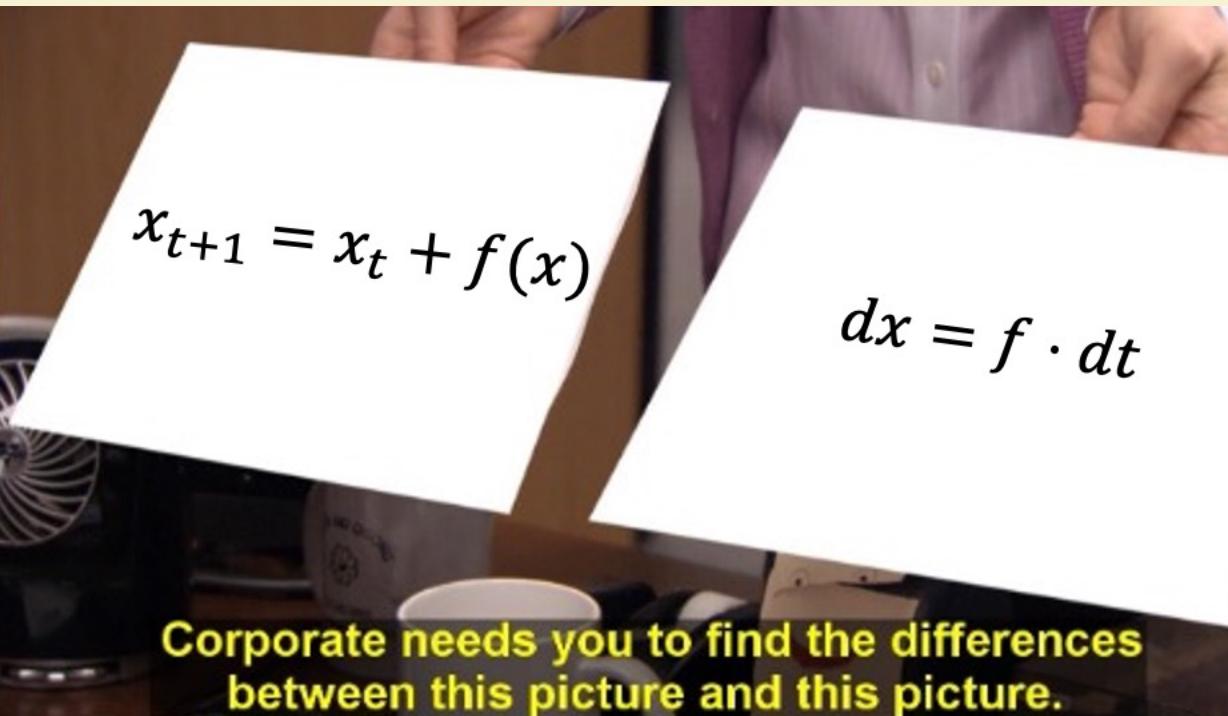
$$f'(y) = \mathbb{E}[2(Y - y) | X = x] = 2(\mathbb{E}[Y | X = x] - y)$$

$$f'(y) = 0 \Leftrightarrow y = \mu$$

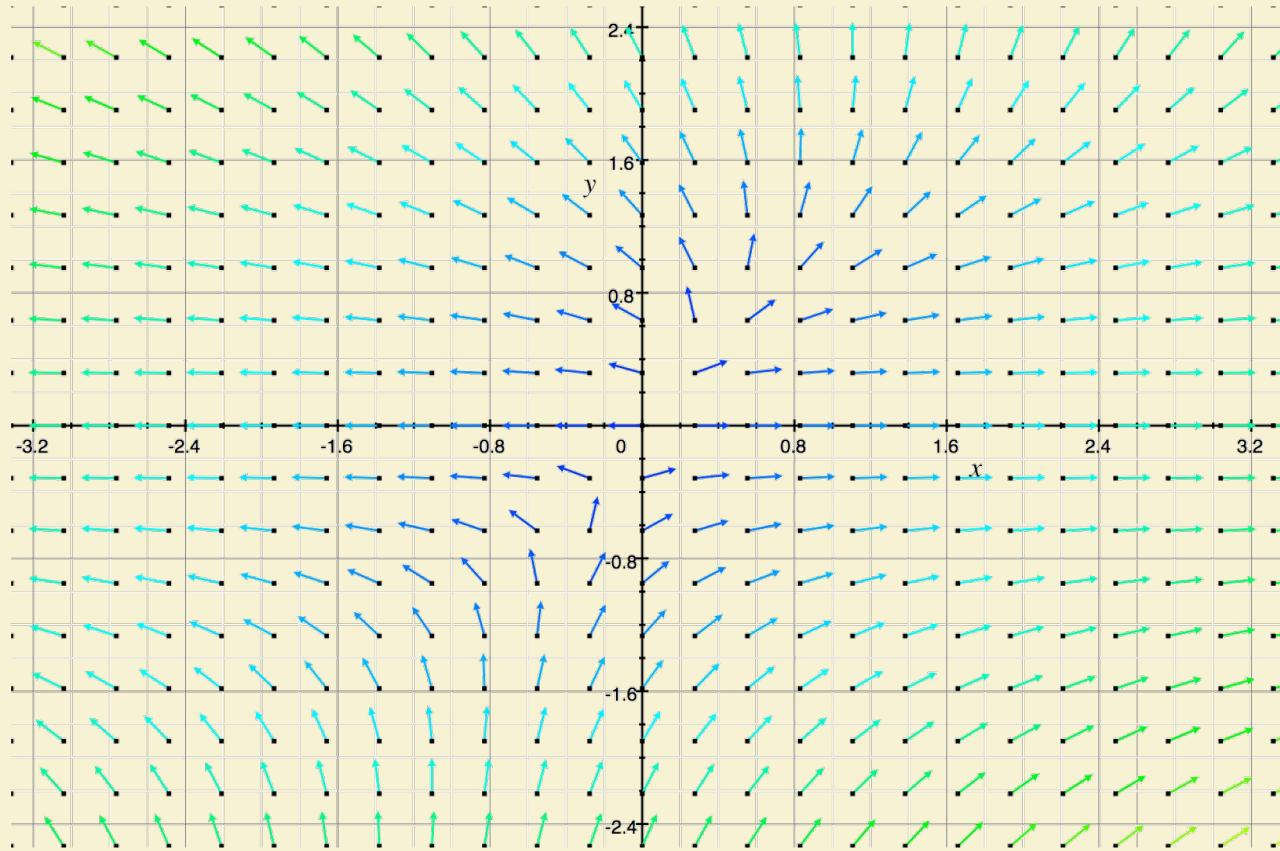


Generalization: In high dimensions  $\mathbb{E}[Y|X = x] = \arg \min \mathbb{E}[\|y - Y\|^2 | X = x]$

# Reminder 4



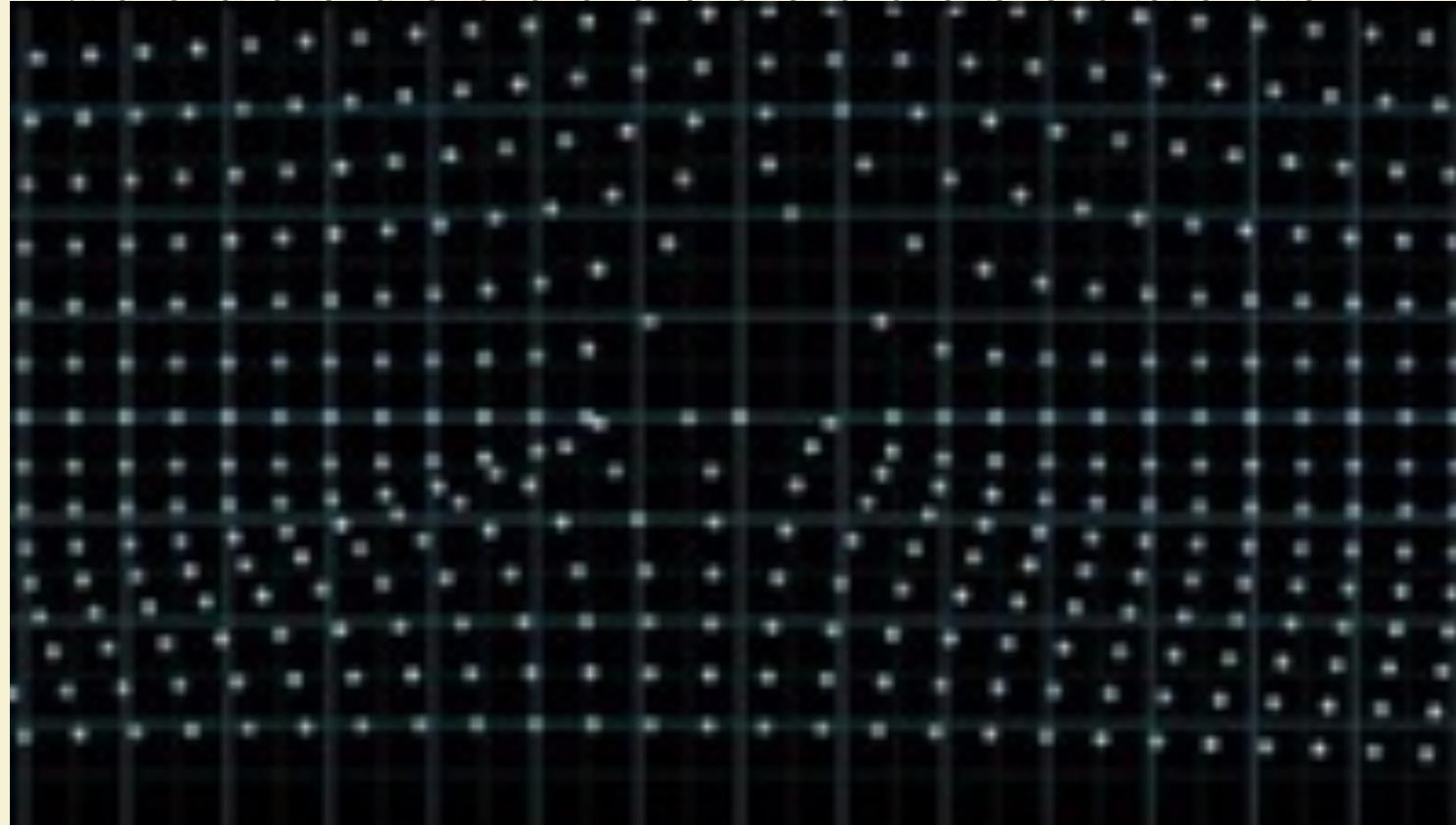
# Digression to divergence



$$F: \mathbb{R}^d \rightarrow \mathbb{R}^d$$

$F(x) = \text{speed \& direction of particles at } x$

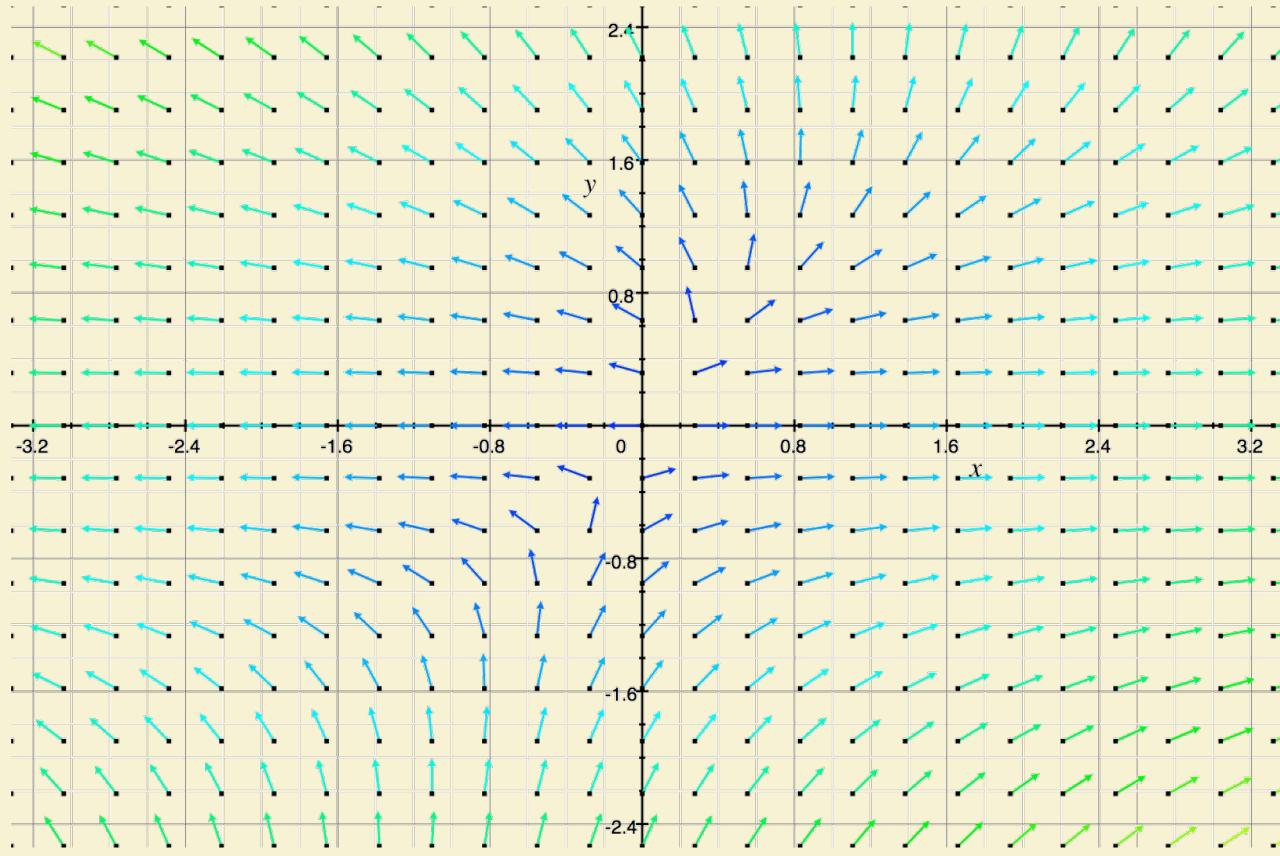
# Digression to divergence



$F: \mathbb{R}^d \rightarrow \mathbb{R}^d$        $F(x)$  = speed & direction of particles at  $x$

$(\nabla \cdot F)(x) \propto$  # particles leaving - # entering at  $x$

# Digression to divergence



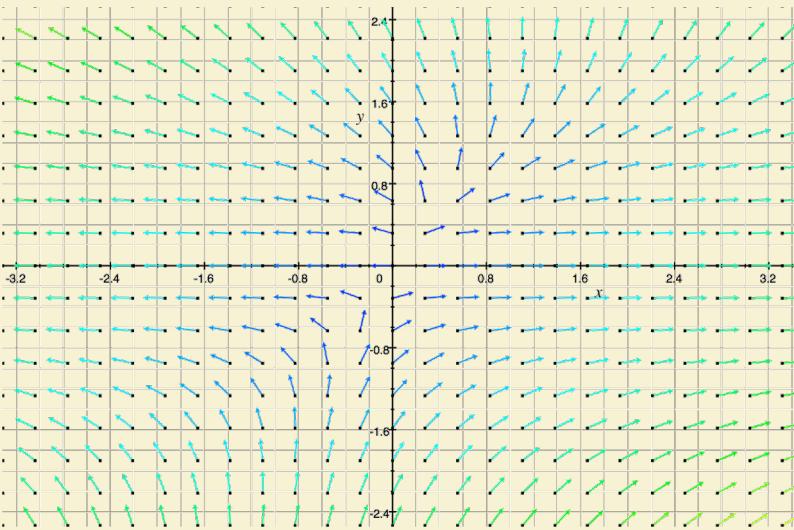
$F: \mathbb{R}^d \rightarrow \mathbb{R}^d$        $F(x)$  = speed & direction of particles at  $x$

$(\nabla \cdot F)(x) \propto$  # particles leaving - # entering at  $x$

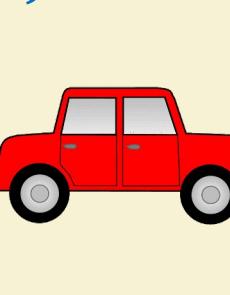
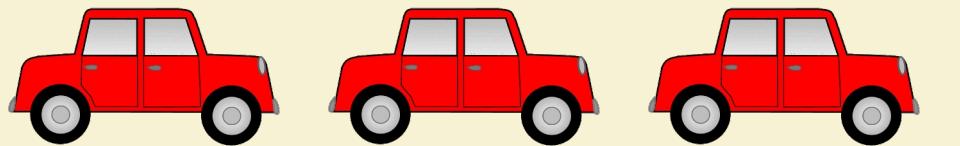
# Digression to divergence

$F: \mathbb{R}^d \rightarrow \mathbb{R}^d$      $F(x)$  = speed & direction of particles at  $x$

$(\nabla \cdot F)(x) \propto$  # particles leaving - # entering at  $x$



Particles at constant speed    i.e.  $F(x) = \text{const}$



$$\#\text{in} = \#\text{out} : \nabla \cdot F = 0$$

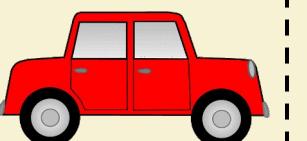


If  $F = \nabla G$

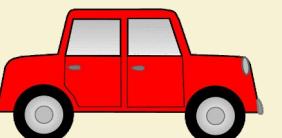
$$\nabla \cdot F = \sum_{i=1}^d \frac{dF(x)_i}{dx_i} = \text{Tr}(\nabla_2 G)$$

Particles accelerating

i.e.  $F'(x) = a$



$$\#\text{out} - \#\text{in} = a$$



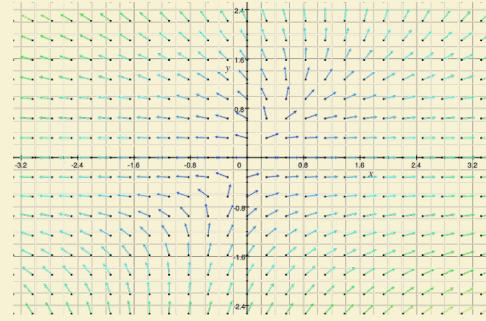
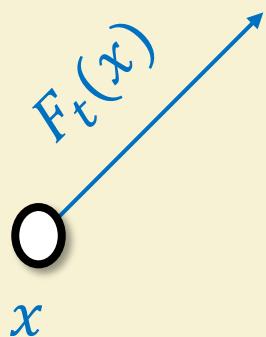
# Evolution of probability distribution

$F: \mathbb{R}^d \rightarrow \mathbb{R}^d$        $F_t(x)$  = speed & direction of particles at  $x$  at time  $t$

$p_t(x)$   
 $(\nabla \cdot F_t)(x) \propto$  # particles leaving - # entering at  $x$  at time  $t$

$$p_{t+dt}(x) = -\nabla \cdot (F_t(x)p_t(x))$$

$$dp_t(x) = -\nabla \cdot (F_t(x)p_t(x)) dt$$



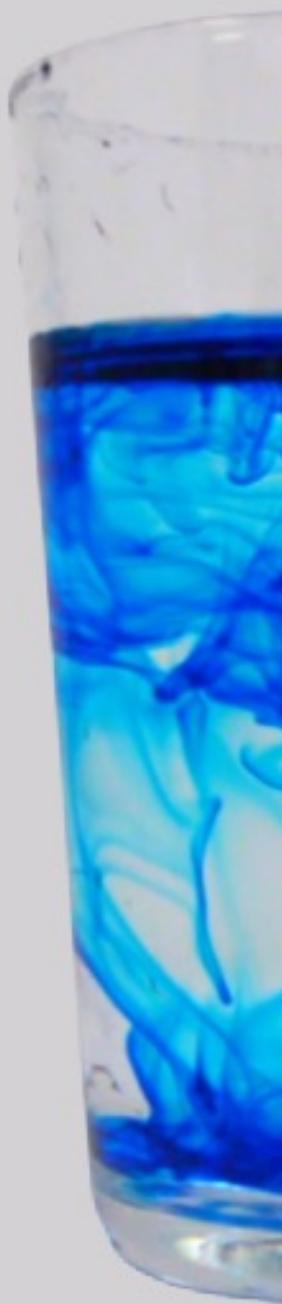
# Diffusion



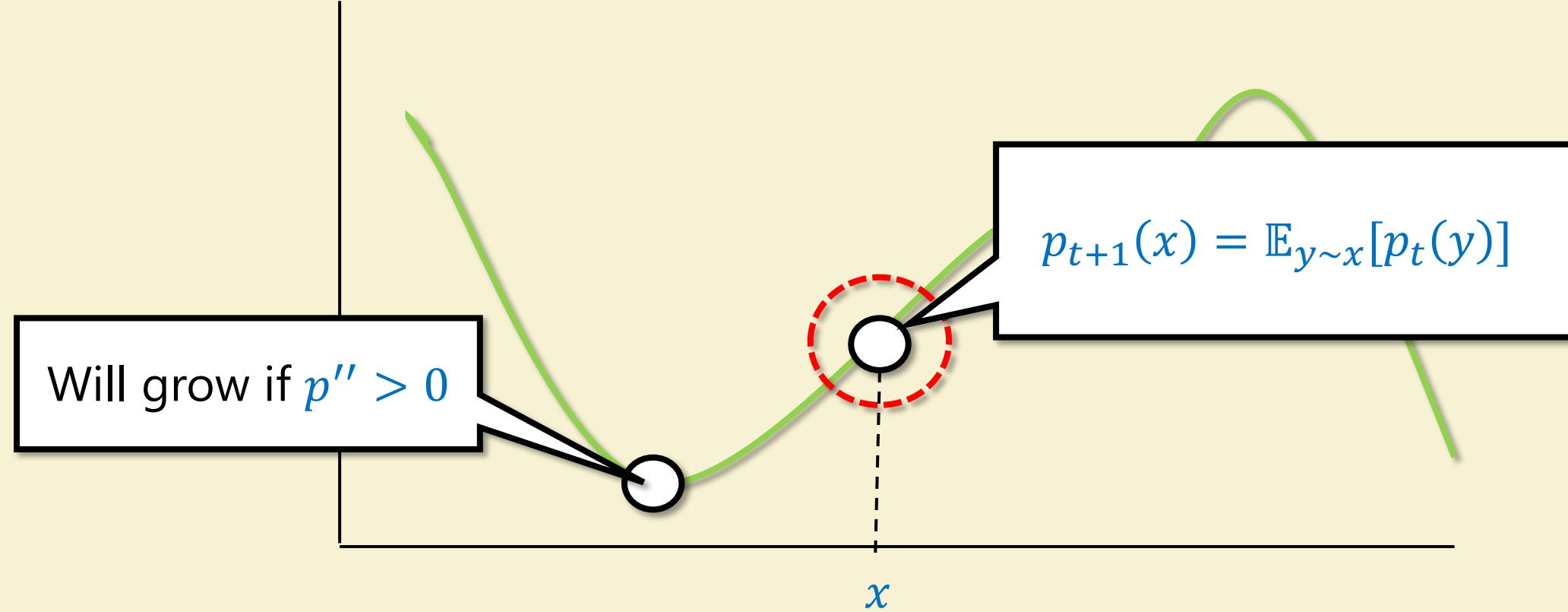
# Diffusion



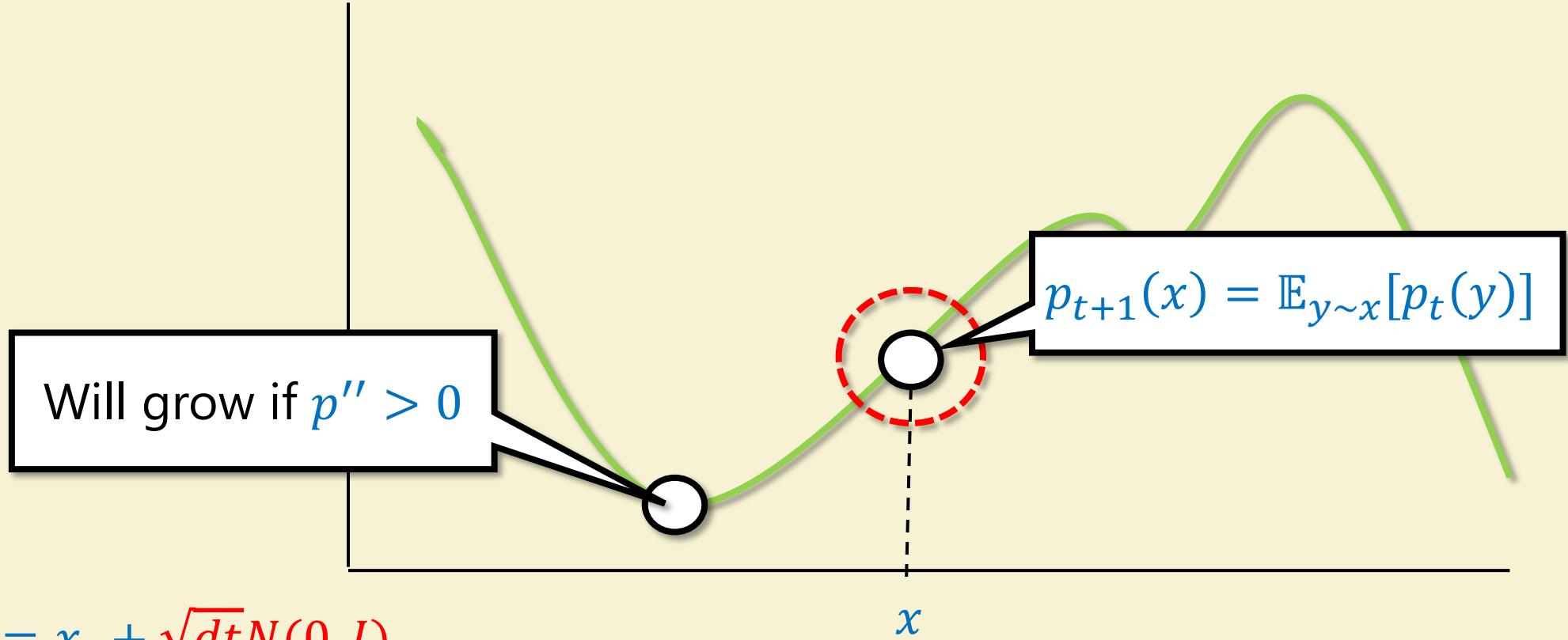
$$p_{t+1}(x) = \mathbb{E}_{y \sim x}[p_t(y)]$$



# Diffusion



# Diffusion



$$p_t(x + \delta) \approx p(x) + \langle \delta, \nabla p(x) \rangle + \frac{\delta^\top \nabla_2 p(x) \delta}{2}$$

$$\mathbb{E} p_t(x + \sqrt{dt}N) \approx p(x) + \sqrt{dt} \mathbb{E} \langle N, \nabla p(x) \rangle + \frac{dt}{2} \mathbb{E} [N^\top \nabla_2 p(x) N] = \frac{1}{2} \text{Tr}(\nabla_2 p) dt$$

Fokker-Plank Equation

$$\frac{dp_t(x)}{dt} = \nabla \cdot \left( \frac{1}{2} \nabla p_t(x) \right)$$

# Take away

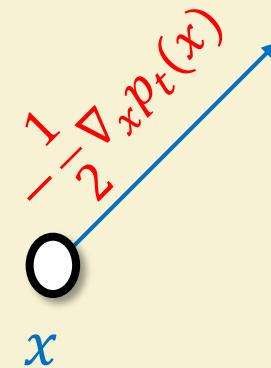
Two views on diffusion:

Stochastic:  $dx = \sqrt{dt} \cdot N(0, I)$

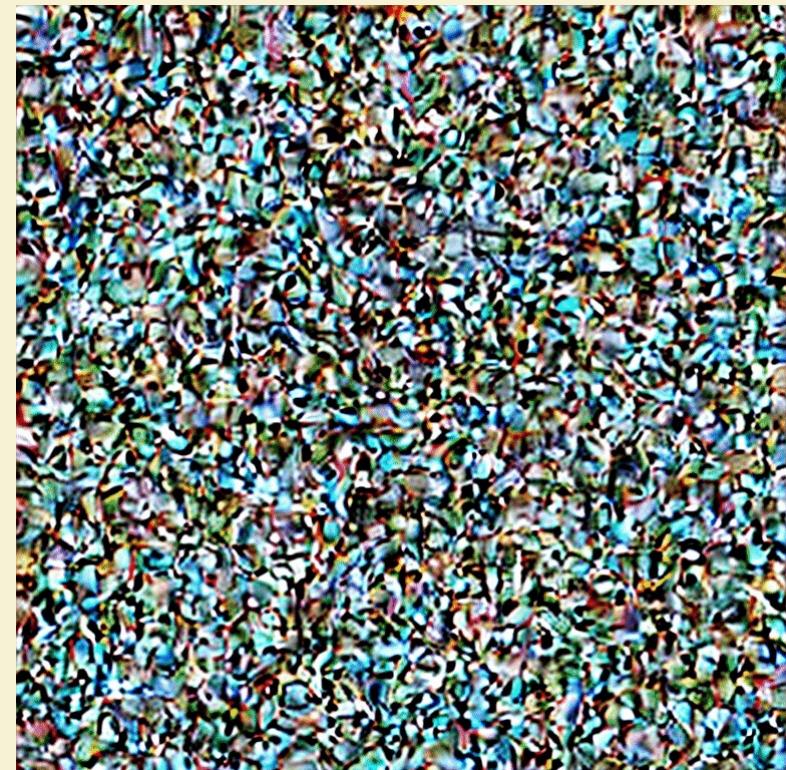
Deterministic:  $\frac{dp_t}{dt} = \frac{1}{2} Tr(\nabla_2 p_t) = \nabla \cdot \frac{1}{2} \nabla p_t = \nabla \cdot \left( p_t \frac{1}{2} \nabla \log p_t \right)$

$$\frac{dx}{dt} = -\frac{1}{2} \nabla \log p_t$$

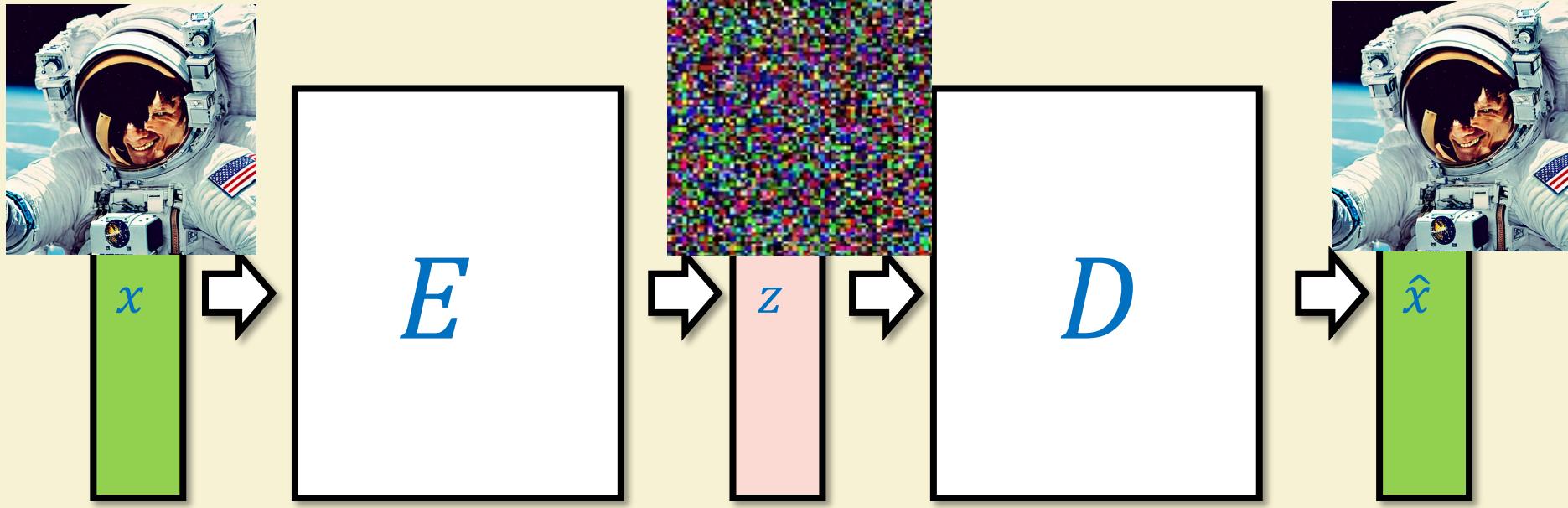
Score matching – note  
connection to EBM



# Diffusion-based Generative models



# Generative Models



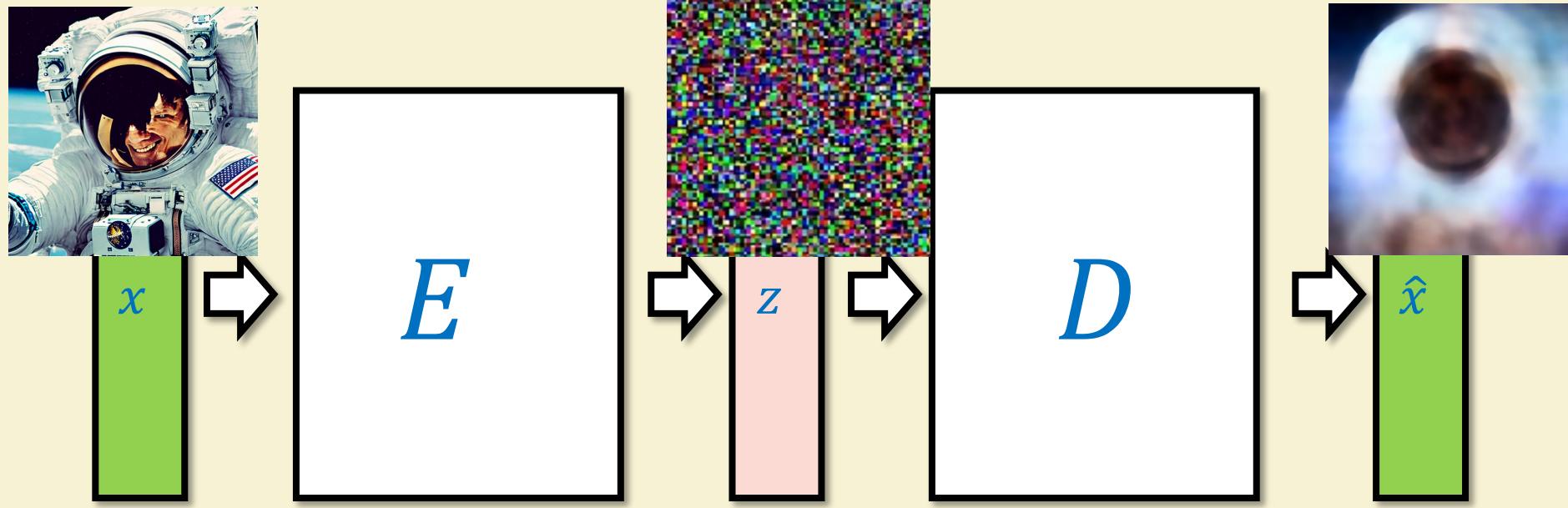
Ideally:

- $D(E(x)) = x$
- $x \sim p \Rightarrow E(x) \sim N(0, I)$

Trivial to achieve:

$$E_\epsilon(x) = \sqrt{\epsilon} \cdot x + N(0, (1 - \epsilon)I)$$

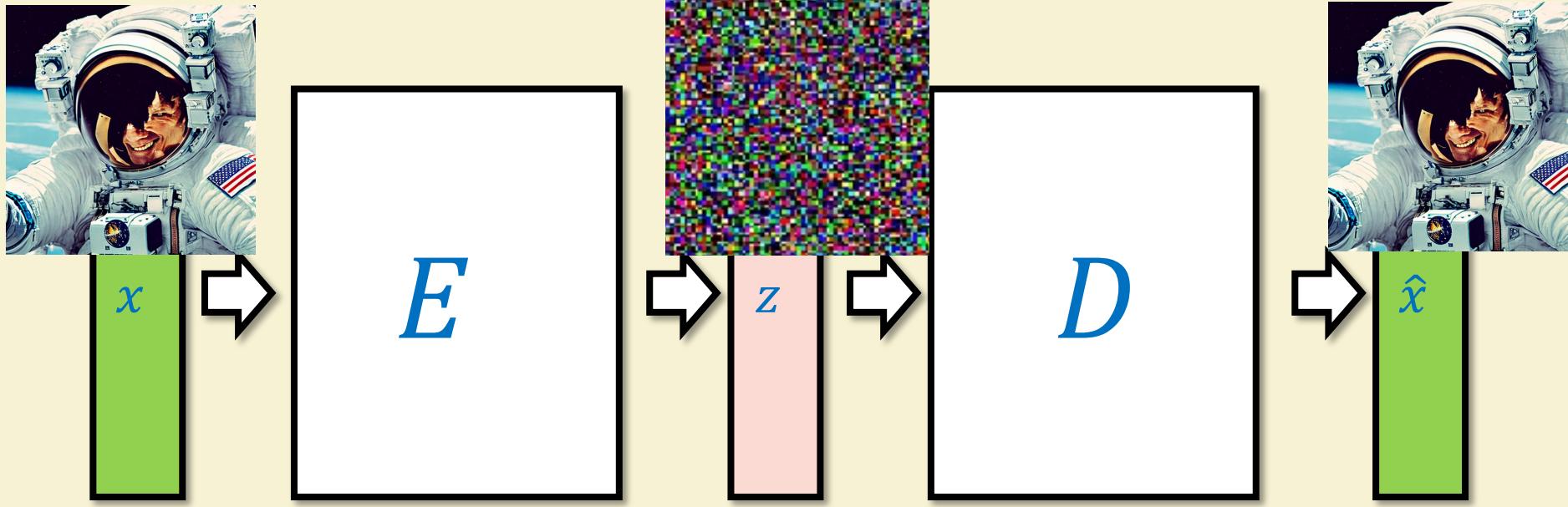
# Diffusion Model v0



Fix  $E(x) = E_\epsilon(x) = \sqrt{\epsilon} \cdot x + N(0, (1 - \epsilon)I)$

$$D_\theta = \arg \min \mathbb{E} \|D_\theta(E_\epsilon(x), \epsilon) - x\|^2$$

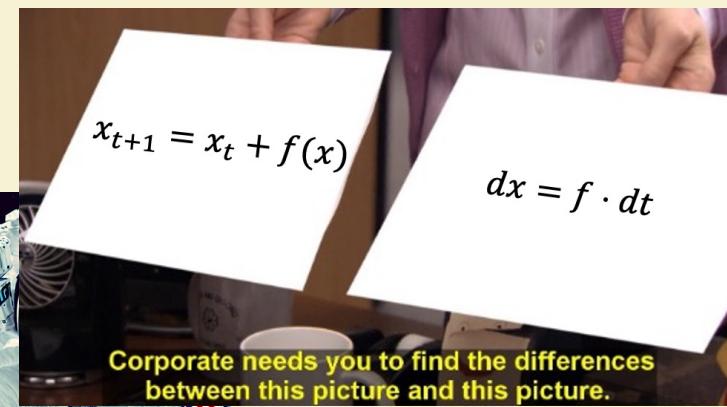
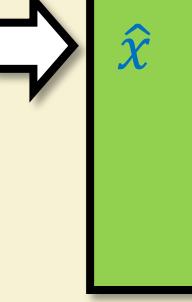
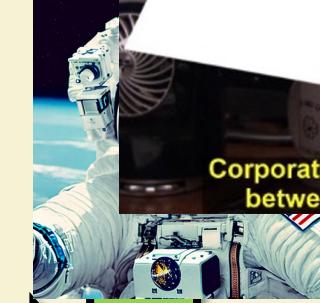
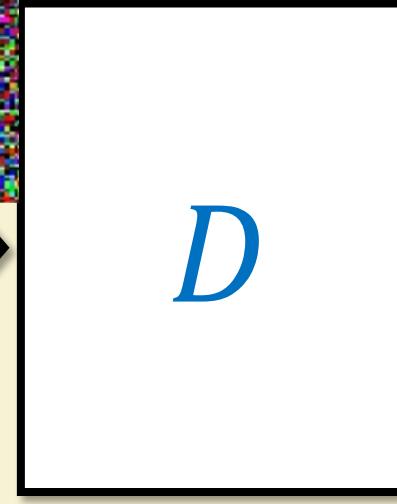
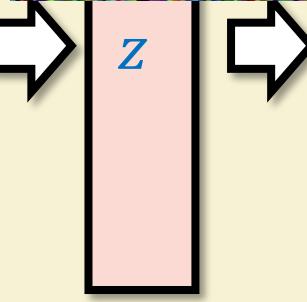
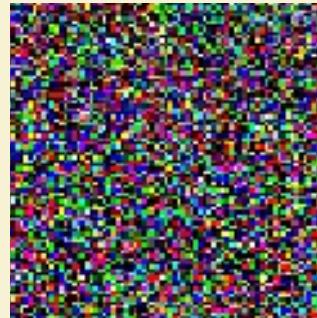
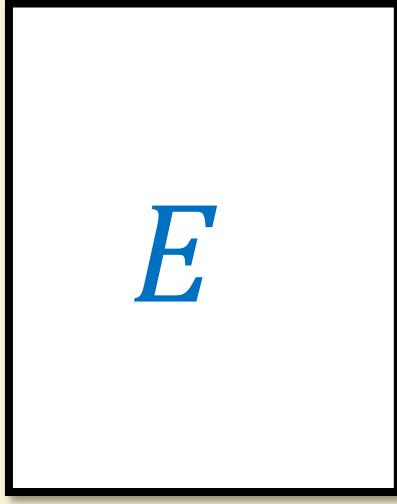
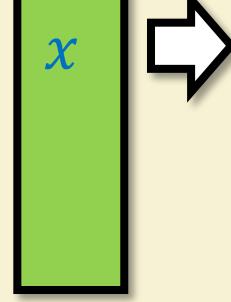
# Diffusion Model v0.5



$$x_0 = x \quad x_{t+1} = \frac{1}{\sqrt{2}} x_t + N(0, \frac{1}{2}I)$$

$$x_n = 2^{-n/2} x_0 + N\left(0, \left(\frac{1}{2} + \frac{1}{2^2} + \dots + \frac{1}{2^n}\right) I\right)$$

# Diffusion Model v1



Discrete version

$$x_0 = x$$

$$x_{t+1} = x_t + N(0, \sigma(t)^2 I)$$

$$x_n = x_0 + N\left(0, \sum_{i=0}^{n-1} \sigma(i)^2 \cdot I\right)$$

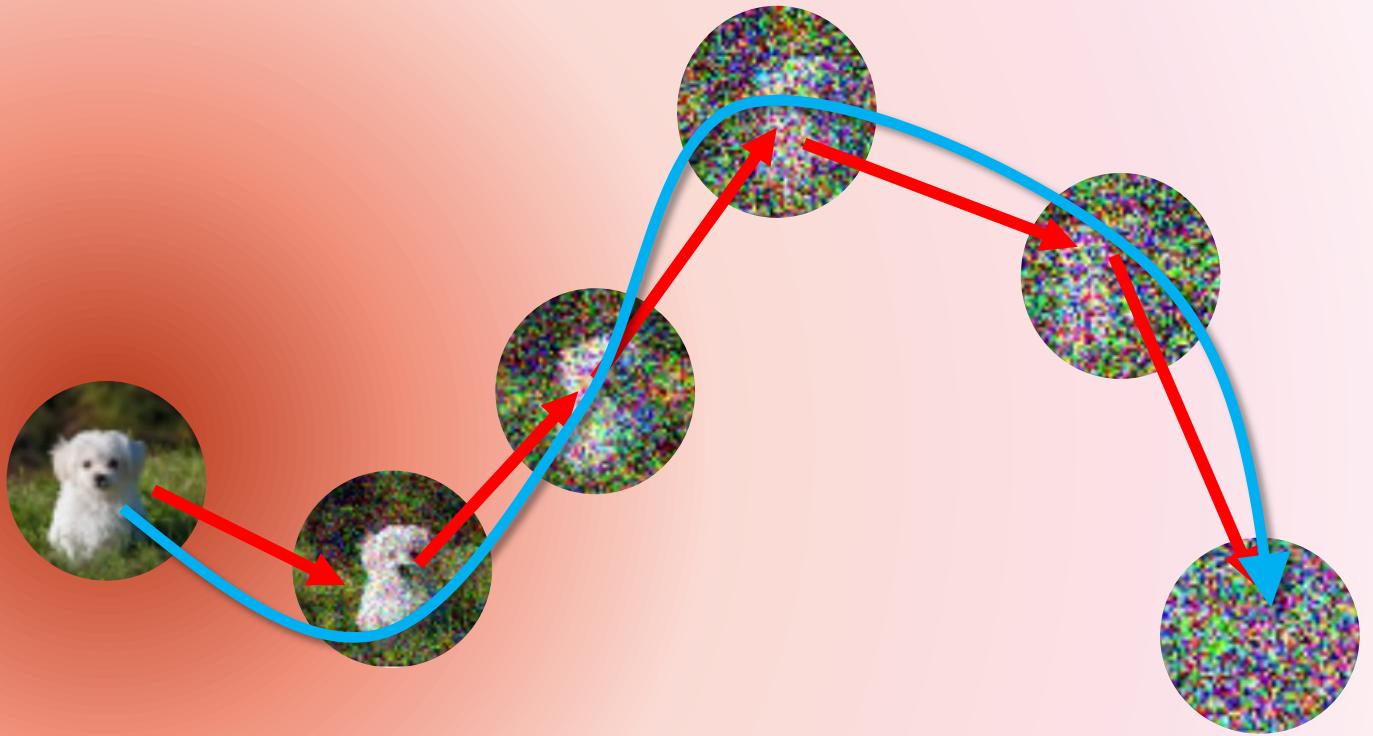
Continuous version

$$x_{t+dt} = x_t + N(0, 2\sigma(t)\sigma'(t)I)dt$$

$$x_n = x_0 + N\left(0, \int_0^n \sigma(t)^2 dt \cdot I\right)$$

$$x_{t+1} = x_t + N(0, \sigma(t)^2 I)$$

$$x_{t+dt} = x_t + N(0, 2\sigma(t)\sigma'(t)I)dt$$



$$x_{t+1} = x_t + N(0, \sigma(t)^2 I)$$

$$x_{t+dt} = x_t + N(0, 2\sigma(t)\sigma'(t)I)dt$$

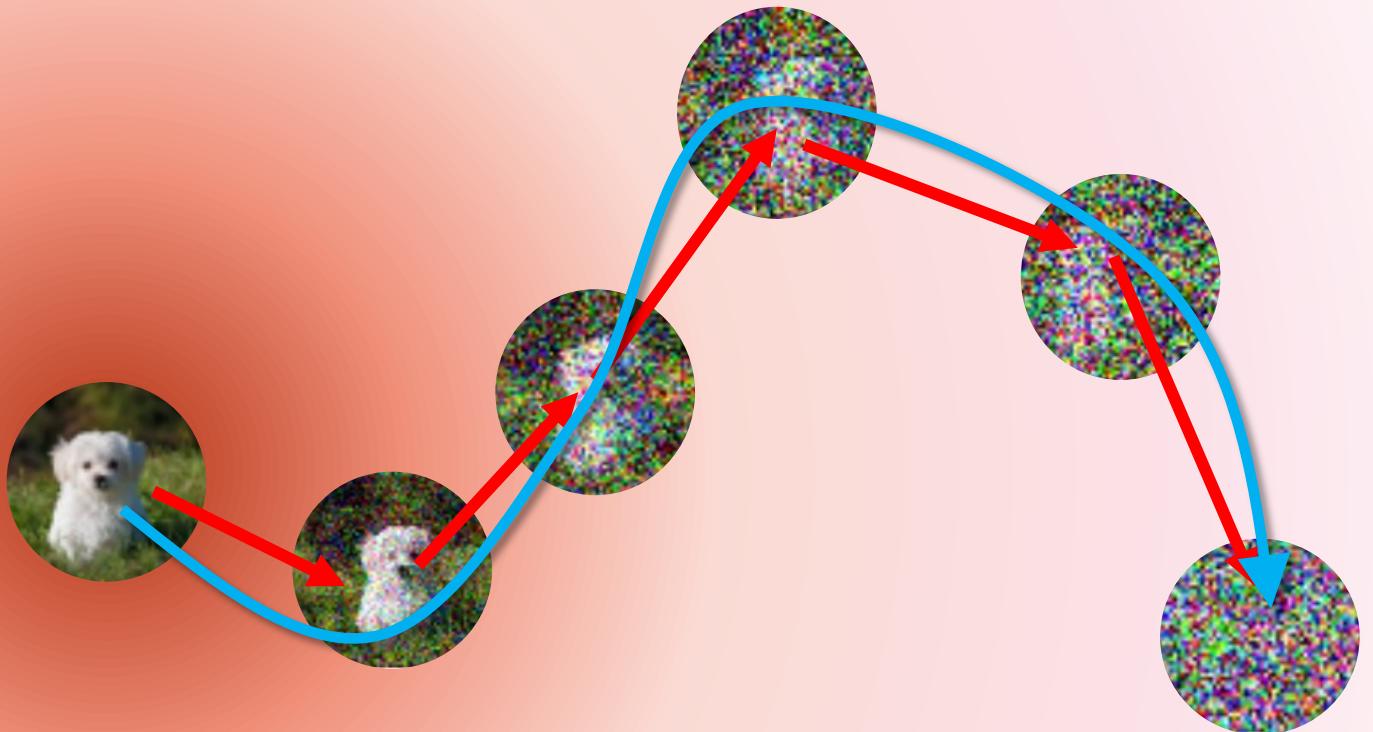
Reversing process:

Deterministic:

Flow from  $x$  in estimated direction of  $x_0$

Randomized:

Sample  $x_{t-1}$  based on  $x_t$  and estimated direction of  $x_0$



# Bayesian approach



Distribution of  $x_t | x_0, x_{t+1}$  is:

1-dim,  $\sigma(t) = \sqrt{t}$  case:  $x_t = N(x_0, t)$        $x_{t+1} = x_t + N(0, 1) = N(x_0, t + 1)$

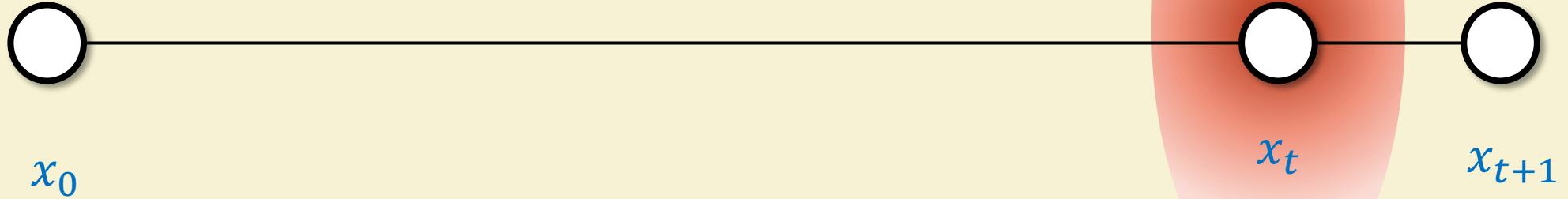
Intuitively:  $x_t | x_0, x_{t+1}$  is  $\approx N(\alpha \cdot x_0 + (1 - \alpha)x_{t+1}, I)$

$$P(x_t = y | x_0, x_{t+1}) \propto \exp(-(y - x_0)^2 / 2t - (x_{t+1} - y)^2 / 2)$$

Mode is when  $\frac{d \log p(y)}{dy} = 0$ :  $-\frac{y}{t} + \frac{x_0}{t} - y + x_{t+1} = 0$        $y \left(1 + \frac{1}{t}\right) = \frac{x_0}{t} + x_{t+1}$

$$y = \frac{1}{t+1} x_0 + \left(1 - \frac{1}{t+1}\right) x_{t+1}$$

# Bayesian approach



Distribution of  $x_t | x_0, x_{t+1}$  is:

1-dim,  $\sigma(t) = \sqrt{t}$  case:  $x_t = N(x_0, t)$        $x_{t+1} = x_t + N(0, 1) = N(x_0, t + 1)$

Intuitively:  $x_t | x_0, x_{t+1}$  is  $\approx N(\alpha \cdot x_0 + (1 - \alpha)x_{t+1}, I)$

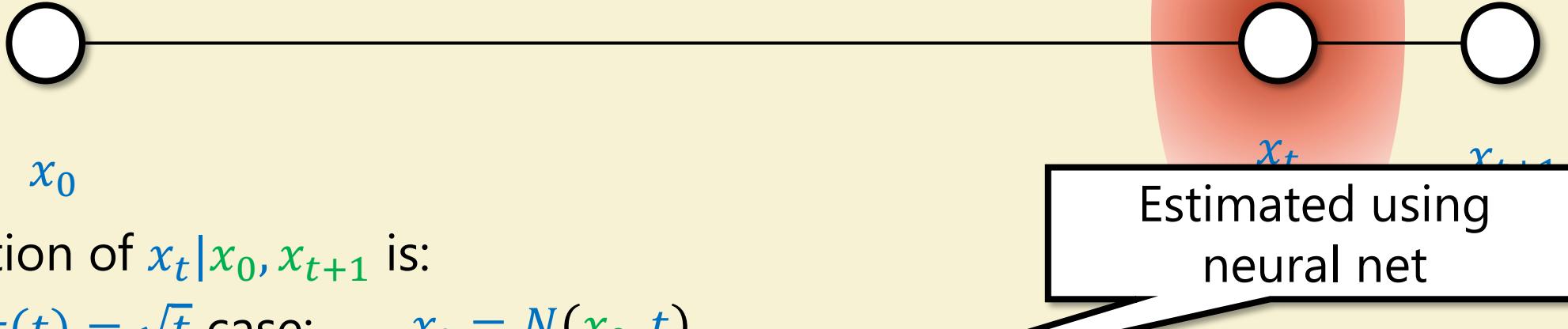
$$y = \frac{1}{t+1} x_0 + \left(1 - \frac{1}{t+1}\right) x_{t+1}$$

Estimated using  
neural net

Reverse  
stochastic process

$$x_t = \frac{1}{t+1} \widehat{x}_0 + \left(1 - \frac{1}{t+1}\right) x_{t+1} + N(0, I)$$

# Bayesian approach



Reverse  
stochastic process

Stochastic  
Differential Equation

$$x_t = \frac{1}{t+1} \widehat{x}_0 + \left(1 - \frac{1}{t+1}\right) x_{t+1} + N(0, I)$$

$$dx_t = -\frac{1}{t} \widehat{x}_0 dt - \frac{1}{t} x_t dt + N(0, I)$$

$$dx_t = -\frac{2\sigma'(t)}{\sigma(t)} \widehat{x}_0 dt - \left(\frac{2\sigma'(t)}{\sigma(t)}\right) x_t dt + N(0, I) \sqrt{2\sigma(t)\sigma'(t)dt}$$

\* For  $\sigma(t) = \sqrt{t}$ ,  $2\sigma(t)\sigma'(t) = 1$ ,  $\frac{\sigma'(t)}{\sigma(t)} = \frac{1}{2t}$

# Deterministic Approach

$$x_{t+1} = x_t + N(0,1) = N(x_0, t+1)$$

$$\frac{dp_t(x)}{dt} = -\nabla \cdot \left( -\frac{1}{2} \nabla p_t(x) \right)$$

$$\frac{dp_t(x)}{dt} = -\nabla \cdot \left( -\frac{1}{2} p_t(x) \nabla \log p_t(x) \right)$$

$$\nabla p = p \cdot \nabla \log p$$

Divergence where particles at  $x$  move at speed  $-\frac{1}{2} \nabla \log p_t(x)$

$$\log p_t(x) = \log \Pr[N(x_0, t) = x] = -\frac{\|x - x_0\|^2}{2t} + \text{const}$$

$$\nabla \log p_t(x) = \log \Pr[N(x_0, t) = x] = \frac{x_0 - x}{t}$$

Estimated using  
neural net

Ordinary  
Differential Equation

$$x_{t-dt} = x_t + \left( \frac{\widehat{x}_0 - x_t}{2t} \right) dt \quad dx = -\sigma'(t)\sigma(t) \left( \frac{\widehat{x}_0 - x_t}{2\sigma^2(t)} \right) dt$$

# Putting it together

Build black box  $D_\theta$  s.t.  $\theta = \arg \min \mathbb{E} \|D_\theta(N(x, \sigma^2 I); \sigma) - x\|^2$

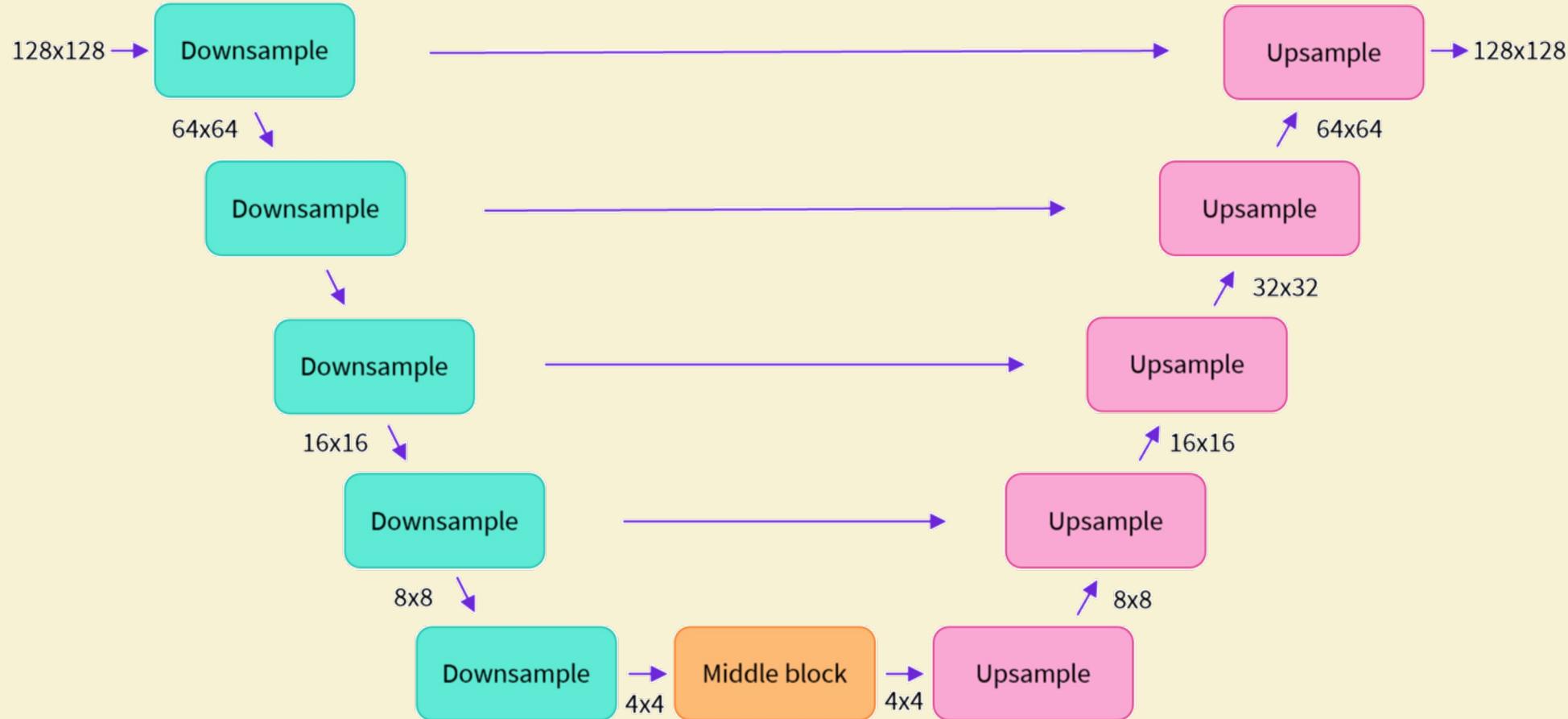
Start with  $x_T = N(0, \sigma_{\max}^2 I)$

and generate  $x_{t_0}, x_{t_1}, \dots, x_{t_k}$  with  $T = t_0 > t_1 > \dots > t_k = 0$

Plugging  $\widehat{x}_0 = D_\theta(x_{t_i}, \sigma(t_i))$  into either SDE or ODE to move from  $t_i$  to  $t_{i+1}$

\* Ignore scaling

# Side note: Architecture: U-Net



$$x_{t+1} = x_t + N(0, \sigma(t)^2 I)$$

$$x_{t+dt} = x_t + N(0, 2\sigma(t)\sigma'(t)I)dt$$

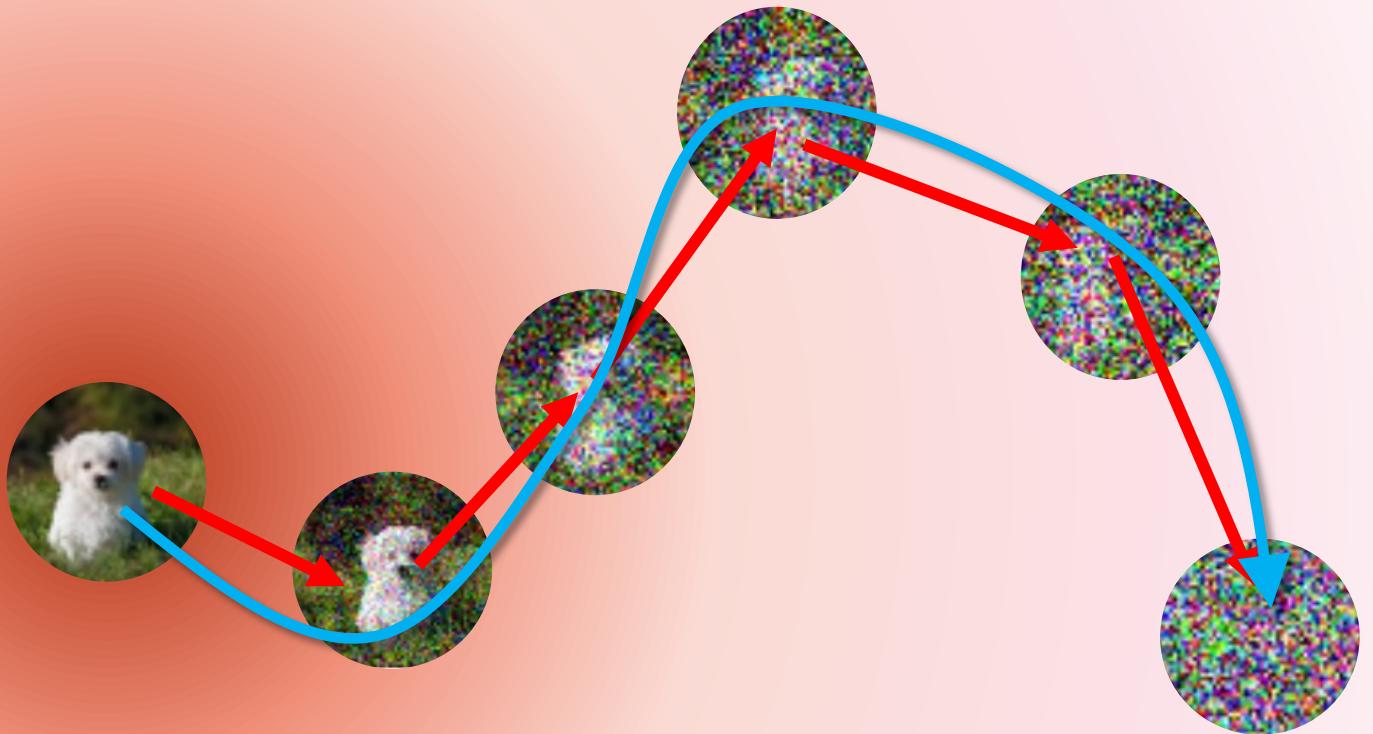
Reversing process:

Deterministic:

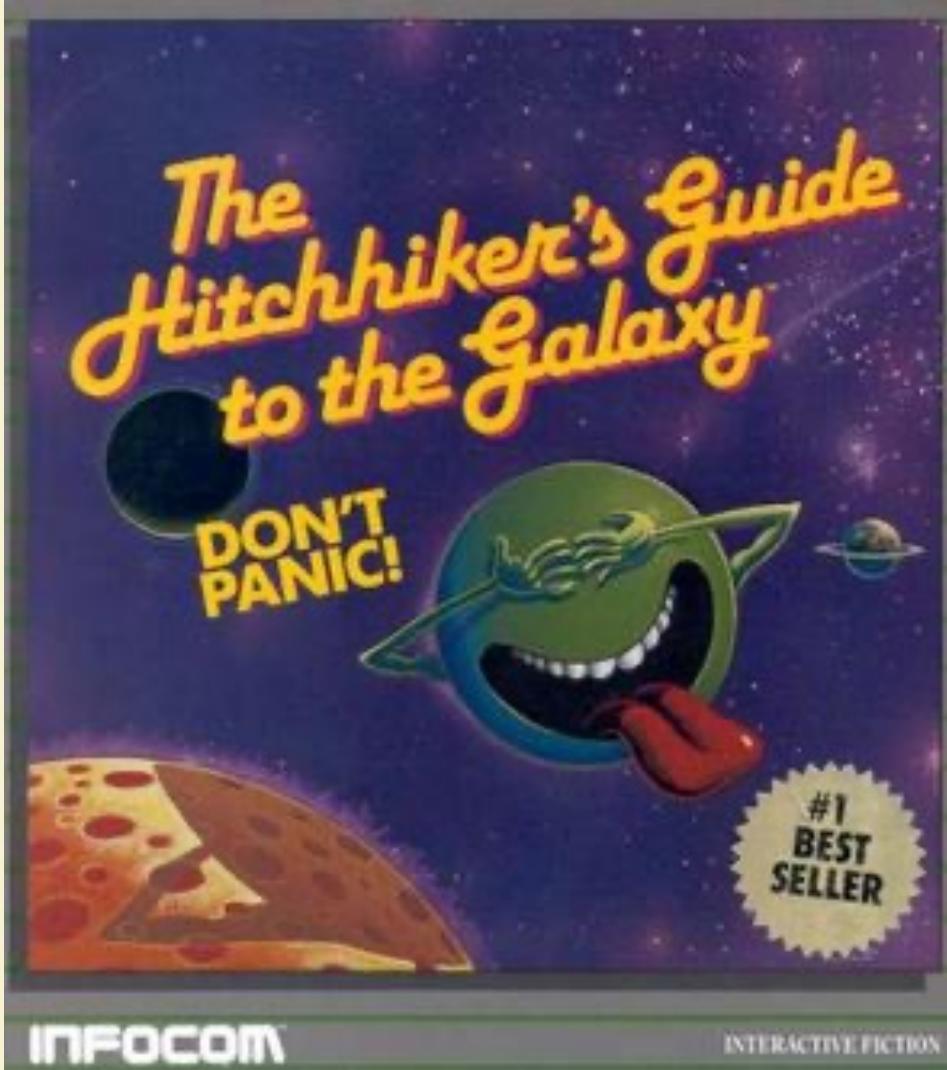
Flow from  $x$  in estimated direction of  $x_0$

Randomized:

Sample  $x_{t-1}$  based on  $x_t$  and estimated direction of  $x_0$



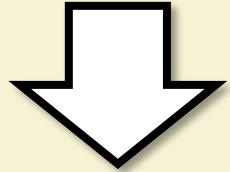
# Guidance



# Classifier Guided Diffusion

Given:  $C: \mathbb{R}^d \times \mathcal{Y} \rightarrow \mathbb{R}$        $C(x + \sigma^2 N(0, I), y; \sigma) = \log \Pr[\text{label}(x) = y]$

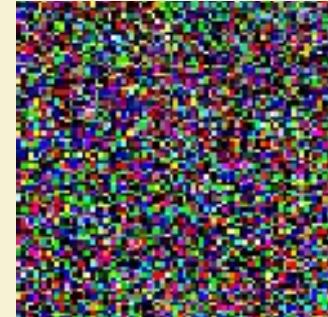
$$dx = -\frac{1}{2} \nabla \log p_t(x)$$



$$dx = -\frac{1}{2} \nabla \log p_t(x) - \lambda \cdot C(x, y; \sqrt{t})$$

# Classifier free guidance

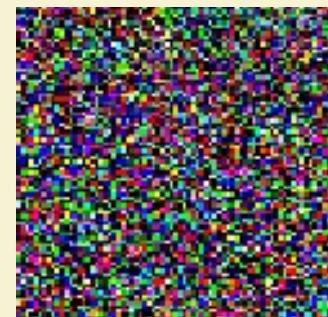
Question 1: What is the unnoised image that led to



?

Question 2: What is the unnoised image that led to

Hint: The label was "astronaut"



?

# Classifier free guidance

Assume  $D_\theta(N(x, \sigma^2 I); \sigma) = \arg \min \mathbb{E} \| \hat{x} - x \|^2 = \mathbb{E}[x | N(x, \sigma^2 I) = z]$

Then  $D_\theta(N(x, \sigma^2 I); \sigma, \textcolor{red}{y}) = \arg \min \mathbb{E} \| \hat{x} - x \|^2 = \mathbb{E}[x | N(x, \sigma^2 I) = z, Y = \textcolor{red}{y}]$

If we provide  $\textcolor{red}{y}$  to denoiser, then signal would be **conditioned** on it.

Choices for  $\textcolor{red}{y}$ :

- Class label
- Textual description (e.g. "alt text") *Mediated through CLIP*

Train  $D_\theta(N(x, \sigma^2 I); \sigma, \textcolor{red}{y}/\emptyset)$

Diffuse via  $(1 + \epsilon)D_\theta(x_t; \sigma(t), \textcolor{red}{y}) - \epsilon \cdot D_\theta(x_t; \sigma(t), \emptyset)$

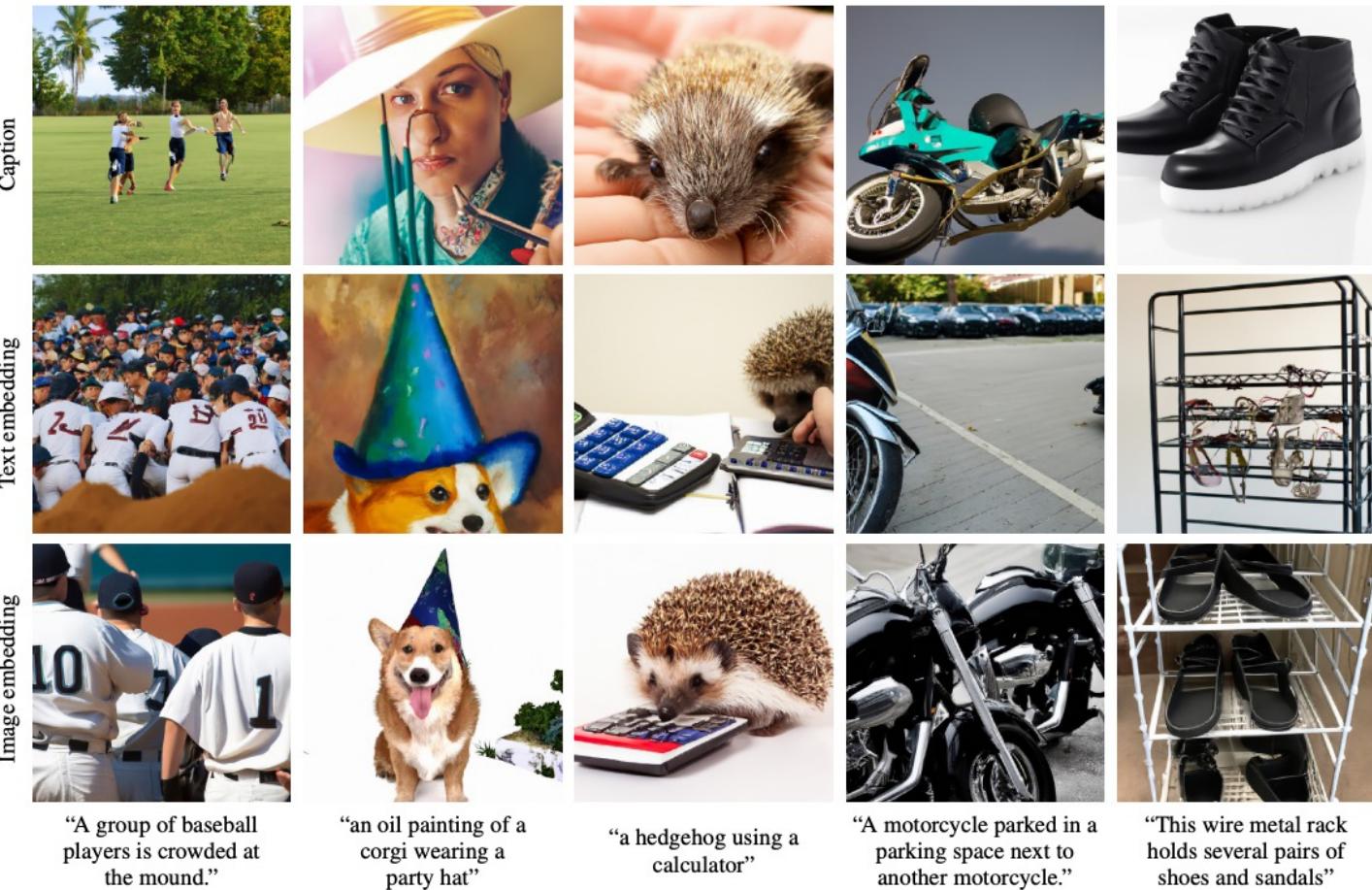


Figure 8: Samples using different conditioning signals for the *same* decoder. In the first row, we pass the text caption to the decoder, and pass a zero vector for the CLIP embedding. In the second row, we pass both the text caption and the CLIP text embedding of the caption. In the third row, we pass the text and a CLIP image embedding generated by an autoregressive prior for the given caption. Note that this decoder is only trained to do the text-to-image generation task (without the CLIP image representation) 5% of the time.



Figure 9: Samples when increasing guidance scale for both unCLIP and GLIDE, using the prompt, “A green vase filled with red roses sitting on top of table.” For unCLIP, we fix the latent vectors sampled from the prior, and only vary the guidance scale of the decoder. For both models, we fix the diffusion noise seed for each column. Samples from unCLIP improve in quality (more realistic lighting and shadows) but do not change in content as we increase guidance scale, preserving semantic diversity even at high decoder guidance scales.

# Measuring Outputs: Quality vs Coverage

## Energy      Entropy

Inception Score:  $C: X \rightarrow \text{Prob}(Y)$  Notation:  $C(\cdot | x)$

$$C(\cdot) = C(\cdot | x), x \sim p_{\text{gen}}$$

$$\log IS(p_{\text{gen}}) := \mathbb{E}_{x \sim p_{\text{gen}}} KL(C(\cdot | x) \| C(\cdot)) = I(C(x); x)$$

$$= H(Y) - H(Y|X)$$

Full as long as one image per label

Zero if model is 100% confident about label

High IS  $\Rightarrow$  High confidence  $\approx$  High “quality”

# Measuring Outputs: Quality vs Coverage

Energy      Entropy

Inception Score:  $C: X \rightarrow \text{Prob}(\mathcal{Y})$

$$\log IS(p_{\text{gen}}) := \mathbb{E}_{x \sim p_{\text{gen}}} KL(C(\cdot | x) \| C(\cdot)) = I(C(x); x)$$

Fréchet Inception Distance:  $C': X \rightarrow \mathbb{R}^d$

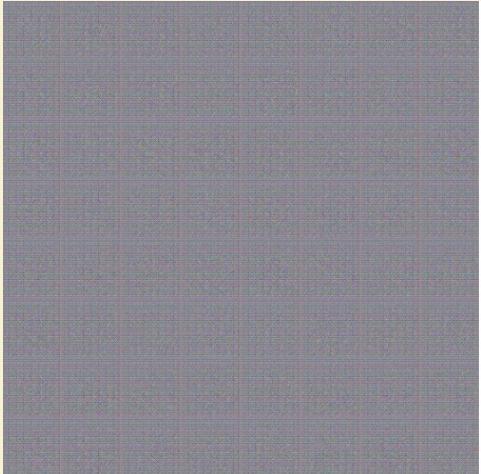
Feature vector

$$FID(p_{\text{gen}}, p_{\text{nat}}) = d^2(\widehat{C'(p_{\text{gen}})}, \widehat{C'(p_{\text{nat}})})$$

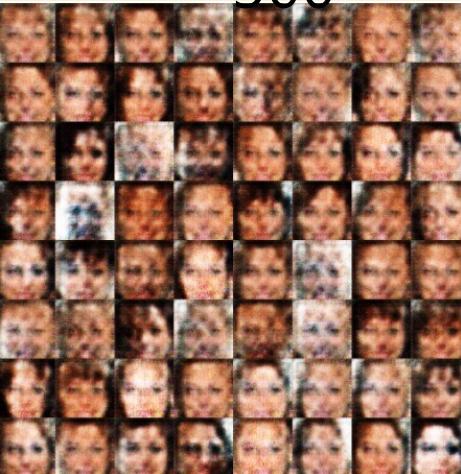
$\widehat{C'(p)}$ : Normal  $N(\mu, \Sigma)$  with  $\mu = \mathbb{E}_{x \sim p}[C'(x)]$  and  $\Sigma = COV_{x \sim p}[C'(x)]$

Recall:  $d^2(N(\mu, \Sigma), N(\mu', \Sigma')) = \|\mu - \mu'\|^2 + \text{tr}(\Sigma + \Sigma' - 2(\Sigma\Sigma')^{1/2})$

500



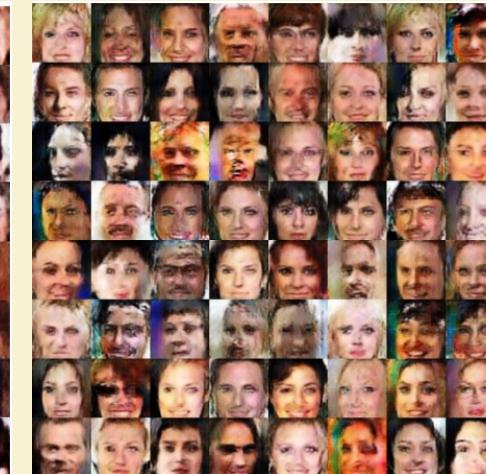
300



100



45



3

